Efficient Multi-View Multi-Target Tracking Using a Distributed Camera Network

Li He, Guoliang Liu, *Member, IEEE*, Guohui Tian, Jianhua Zhang, Ze Ji

***Abstract*— In this paper, we propose a multi-target tracking method using a distributed camera network, which can effectively handle the occlusion and reidenfication problems by combining advanced deep learning and distributed information fusion. The targets are first detected using a fast object detection method based on deep learning. We then combine the deep visual feature information and spatial trajectory information in the Hungarian algorithm for robust targets association. The deep visual feature information is extracted from a convolutional neural network, which is pre-trained using a large-scale person reidentification dataset. The spatial trajectories of multiple targets in our framework are derived from a multiple view information fusion method, which employs an information weighted consensus filter for fusion and tracking. In addition, we also propose an efficient track processing method for ID assignment using multiple view information. The experiments on public datasets show that the proposed method is robust to solve the occlusion problem and reidentification problem, and can achieve superior performance compared to the state of the art methods.**

***Index Terms*—Multi-target tracking, distributed camera network, information fusion, data association, SORT**

I. Introduction

MULTIPLE object tracking (MOT) has many applications nowadays, e.g., surveillance, monitoring, crowd behavior analysis, etc. However, MOT is still a challenging task since it needs to simultaneously solve the object detection, trajectory estimation, data association and reidentification problems. To detect the object, various sensors can be employed according to the requirements of the specific task, e.g., radar, laser, camera, sonar, etc. Compared to other sensors, cameras are similar to human eyes that can capture colorful information about the targets, whereas radar, laser and sonar only measure distance information. Therefore, cameras can be more useful for object detection, association and reidentification, since visual features are more robust than the position information for reducing ambiguity.

For visual multiple target tracking, robust visual feature extraction from targets forms the core capability to handle the varying ambient lighting conditions and target poses. Recently, deep learning based object detection and recognition techniques become popular due to its real time performance, high accuracy and robustness. Therefore, we here adopt the advanced deep neural network YOLOv3 for fast object detection [1], and extract visual features from a pre-trained convolutional neural network (CNN), which is similar to the framework of Deep SORT [2]. Our experiments show that the deep learning based object detection and feature representation methods are more robust than the classical visual features.

Another challenging problem of the visual MOT task is the occlusion, i.e., the target can be occluded by other objects or out of the current field of view (FOV), which can be solved by multiple view information fusion. Therefore, we here propose a distributed multiple view fusion method using a novel information consensus filter (ICF) for robust trajectory tracking. In contrast to the centralized network structure, the distributed camera network has no central fusion node, and each camera only communicates with the neighbor camera as shown in Fig.1. In this way, the distributed network can be

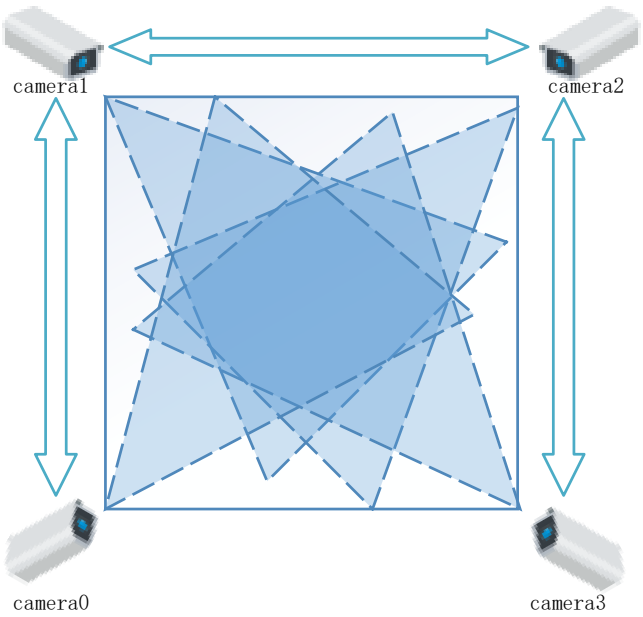


Fig. 1: We use a distributed camera network to demonstrate the proposed idea. The cameras share overlapping areas, such that the targets can be observed from multiple view. Each camera can communicate with neighbor cameras, and exchange information for fusion.

This research was supported by the National Key R&D Program of China (2018YFB1306500), National Natural Science Foundation of China (61603213, 91748115), Young Scholars Program of Shandong University (2018WLJH71), Hebei Provincial Natural Science Foundation (F2017202062), the Fundamental Research Funds of Shandong University, and the Taishan Scholars Program of Shandong Province. (Corresponding author: Guoliang Liu).

L. He, G. Liu, and G. Tian are with the School of Control Science and Engineering, Shandong University, Jinan, 250061 China (e-mail:549452210@qq.com, liuguoliang@sdu.edu.cn, g.h.tian@sdu.edu.cn).

J. Zhang is with School of Mechanical Engineering, Hebei University of Technology, Tianjin, 300131 China (jhzhang@hebut.edu.cn).

Z. Ji is with the Robotics and Autonomous Systems Laboratory, School of Engineering, Cardiff University, Cardiff, CF10 3AT UK (jiz1@cardiff.ac.uk).

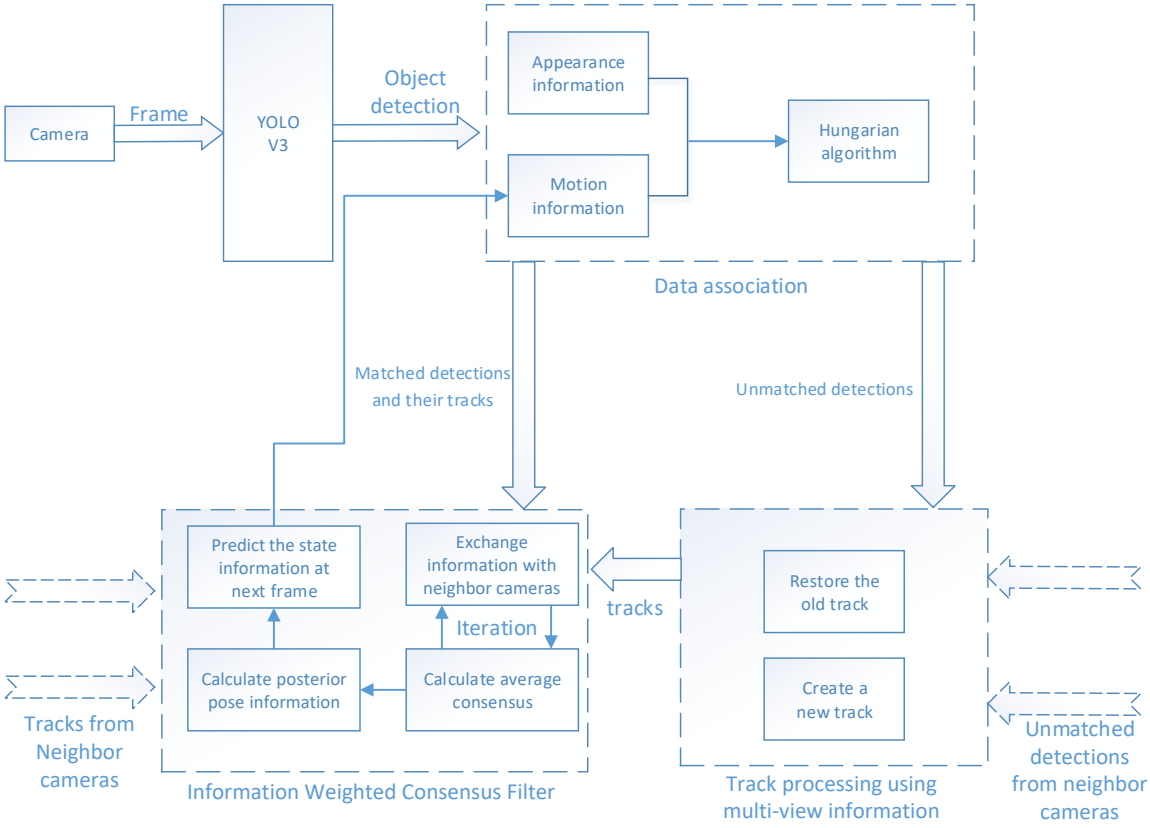


Fig. 2: Overall architecture of our distributed multi-view multi-target tracking system.

more scalable, robust to sensor failures and can save network bandwidth, such that it has a great potential for solving the MOT problem of a larger area.

The architecture of the proposed distributed multiple target tracking method can be seen in Fig.2. The targets are first detected using YOLOv3, and then the visual appearance features of the targets are extracted using a pre-trained neural network. From current view, the detected targets are then associated to previous tracked targets for ID assignment using the Hungarian algorithm that employs both visual appearance information and predicted motion information from ICF filter. If the association is successful, the current target information is used to update the state of the ICF filter [3]. If the association is failed, then the unmatched targets are sent to neighbor cameras for cross view ID assignment.

The main contribution of our work is twofold. First, we propose a distributed multi-view multi-target tracking method that can benefit from both the deep appearance visual features and distributed trajectory estimation, which can be seen as a distributed extension of the state of the art work Deep SORT. Second, we propose a distributed track management method for restoring old tracks or creating new tracks for unmatched targets. In this way, our method is more robust to deal with the object reidentification problem and occlusion problem compared with the original Deep SORT method. To demonstrate the proposed idea, we compare our work with the state of the art methods on the public EPFL datasets. The experimental results show that our method can achieve higher tracking accuracy.

II. Related Work

A. Multi-Target Tracking

Multi-target tracking has been extensively studied due to its wide range of applications. Multiple hypothesis tracking (MHT) and joint probability data association filter (JPDAF) are two popular traditional methods for this task. The JPDAF associates observed measurements of current frame with existing targets using a joint probabilistic value, which suffers from combinatorial complexity problem since it matches all possible assignments of measurements to existed targets to compute the joint probabilistic values [4]. The MHT maintains all potential track hypothesis for each existed target, which is slow and memory intensive for visual object tracking problem [5].

Recently, with the development of deep learning techniques for object detection, simple online and realtime tracking (SORT) [6] is proposed for fast and effective multiple targets tracking, which uses a Kalman filter for object state estimation, and employ the novel Hungarian algorithm for data association. To further improve the SORT algorithm using deep learning ideas, Deep SORT is proposed in [2] that combines the visual appearance features for ID assignment and reidentification, The visual appearance features are extracted from a

pre-trained convolutional neural network (CNN). The Deep SORT shows better performance for object reidentification than the original SORT.

B. Multi-View Tracking

Multi-view tracking can be useful for solving occlusion problem and improving track accuracy, since it can get relatively complete information about the targets. For fusing the multiple view information from camera networks, a centralized or distributed network structure can be used. Fleuret *et al.* [7] and Berclaz *et al.* [8] propose a probabilistic occupancy map for tracking multiple targets using multiple view information. The data association problem can be solved using a dynamic programming or linear programming method. Xu *et al.* [9] compose multiple cues with proper scheduling by a compositional structure optimization method, such that different cue can play a key role for different situation. Tang *et al.* [10] use an energy minimization method to solve data association problem by combining the visual and semantic attributes. He *et al.* [11] employ a neural network to transfer partial view of the targets to real top view, such that the image of objects from different views can be transferred to the common top view. These methods use a centralized network structure for data communication and fusion. In contrast, Kamal *et al.* [12] propose a distributed multi-target information consensus (MTIC) method using information weighted consensus filter (ICF) for multiple view fusion and JPDAF for data association. The camera nodes can communicate with neighbor cameras for information exchanging and fusion. There is no central sensor node, such that the distributed solution is more robust to sensor failure problem, and scalability to large scale network.

III. METHOD

Overall architecture of the proposed distributed multi-view multi-target tracking system can be seen in the Fig.2. We first use the novel YOLOv3 to detect targets for each camera, which can extract a rectangle bounding box of the detected target with a high frame rate. The visual appearance information of the target is then derived from a pre-trained convolution neural network. We combine the visual appearance feature and location information of the target for data association using the novel Hungarian algorithm. The location information of the target can be fused from multiple view information using an information weighted consensus filter.

A. Target Detection

Target detection refers to acquire different objects in an image and determine their type and location. With the rapid development of deep learning in the direction of target detection, the target detection method based on deep learning is very robust for complex environment with illumination change and occlusion problems, which mainly have two research directions: two-stage method and one-stage method [13]. The two-stage method first predicts a number of candidate frames that may have targets, then resizes and classifies the frames to have the precise location, size and category of the targets, e.g., Faster R-CNN [14]. The One-stage method omits the first step and directly predicts the location and category of the target, e.g., YOLOv3. Compared to the two-stage method, the one-stage method is normally faster with comparable performance. Therefore, we here use YOLOv3 as our target detector.

B. Data Association

We use a simple Hungarian algorithm for data association, which employs visual appearance information and target position information. The visual appearance information is derived from a pre-trained convolution neural network as shown in [2], whereas the target position information is a projected coordinate on the ground plane which can be derived from image plane using a calibrated homography matrix [7].

For target position information, we calculate the Mahalanobis distance between the measured position of current view and the predicted trajectory from last time step:

$$d^{(1)}(i,j) = (\boldsymbol{m}_j - \boldsymbol{y}_i)^T \boldsymbol{S}_i^{-1} (\boldsymbol{m}_j - \boldsymbol{y}_i), \tag{1}$$

where i, j represent the i-th trajectory and the j-th measurement, respectively. $\boldsymbol{y}$ and $\boldsymbol{S}$ represent the prediction value and prediction covariance of the trajectory from the information consensus filter (ICF). $\boldsymbol{m}$ represents the new measured position value. In addition, we use a threshold function to omit the unrelated candidates:

$$b_{i,j}^{(1)} = 1, \ if \ d^{(1)}(i,j) \leq t^{(1)}, \tag{2}$$

which means the association between the i-th track and the j-th probe is admissible if the Mahalanobis distance d is less than t, i.e., b is set to 1.

For the appearance information, we store a number of feature vectors of the associated target in list $\mathcal{R}_i$ to handle appearance changing. For new measurements, we calculate the minimum cosine distance of the 128-dimensional feature vector of the target with existed tracks as

$$d^{(2)}(i,j) = \min\{1 - \boldsymbol{r}_j^T \boldsymbol{r}_k^{(i)} | r_k^{(i)} \in \mathcal{R}_i\}, \tag{3}$$

where $\boldsymbol{r}_j$ is the feature vector of the j-th target in current view, $\boldsymbol{r}_k^{(i)}$ is the feature vector of the i-th existed target with k-th descriptor in $\mathcal{R}_i$. Again, we introduce a threshold function to indicate whether the association is possible:

$$b_{i,j}^{(2)} = 1, \ if \ d^{(2)}(i,j) \leq t^{(2)}. \tag{4}$$

To combine the two metrics, we use a weighted sum:

$$c_{i,j} = \lambda d^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j), \tag{5}$$

where λ is an adjustable weight parameter. If and only if both metrics $b_{i,j}^{(1)}$ and $b_{i,j}^{(2)}$ are within the gating range, they are allowed to be associated. A matrix whose element is $c_{i,j}$ can be used as the input of the Hungarian algorithm [15] that can find maximum-weight matchings in bipartite graphs, which is also known as the assignment problem.

Algorithm 1 Track Processing Using Multi-View Information

Input: D_i: unmatched targets of i-th view.

1: **for** $D \in D_i$ **do**
2: Match with existed tracks and unmatched targets in other views using Euclidean distance
3: **if** Match successful with existed tracks **then**
4: Assign the same ID as the existed track to this target
5: **else if** Match successful with unmatched targets **then**
6: Assign a new ID for these unmatched targets
7: **end if**
8: **end for**

C. Track Processing Using Multi-View Information

Track processing step is mainly for target ID management, e.g., restore old tracks or create new tracks, which can be seen in Algorithm 1. If an unmatched target is found in current view, we first match its position information to tracks in other views using Euclidean distance. If one matched candidate is found, then this unmatched target get the same ID as this matched candidate from other views. If the match process is failed, we then wait for a few frames to check whether it is a real target. For new targets, we also check whether this target shows up in other views. If other views can see this new target, we then initialize a new ID for this target. We remove these tracks that have been disappeared for more than 30 seconds in current view.

D. Information Weighted Consensus Filter for Multi-View Information Fusion

Information weighted consensus filter (ICF) is an effective distributed state estimation method. We here use the ICF for multiple view information fusion to estimate the position of targets. The ICF mainly has three steps: state prediction, measurement update and weighted consensus. For state prediction, we here use a linear constant velocity model to predict the state vector x and information matrix W of i-th camera at the next time step:

$$W_i^-(t) = (\Phi(W_i^+(t-1))^{-1}\Phi^T + Q)^{-1}, \tag{6}$$

$$x_i^-(t) = \Phi x_i^+(t-1), \tag{7}$$

where Φ is a linear state transition matrix and Q is the process noise covariance. The predicted position information is sent to the association module to find the matched measurements. For measurement updating, we first calculate the information vector v_i and information matrix V_i using the current measurement z_i:

$$V_i = \frac{1}{N}W_i^- + H_i^T R_i H_i, \tag{8}$$

$$v_i = \frac{1}{N}W_i^- x_i^- + H_i^T R_i z_i, \tag{9}$$

where $x_i^-, W_i^-, H_i, R_i, N$ are prior state vector, information matrix, observation matrix, measurement noise covariance and the number of cameras respectively. For weighted consensus, the i-th camera exchanges its information vector v_i and information matrix V_i with neighbor connected camera nodes $\mathcal{N}_i$. The weighted average consensus at k-th iteration is performed according to :

$$V_i^k = V_i^{k-1} + \epsilon \sum_{j \in \mathcal{N}_i} (V_j^{k-1} - V_i^{k-1}), \tag{10}$$

$$v_i^k = v_i^{k-1} + \epsilon \sum_{j \in \mathcal{N}_i} (v_j^{k-1} - v_i^{k-1}). \tag{11}$$

The consensus can be performed for a number of iterations until the filter is converged or the maximum number of iterations is achieved. Finally, the posterior state vector $x_i^+(t)$ and information matrix $W_i^+(t)$ at current time step is derived as

$$x_i^+(t) = (V_i^K)^{-1} v_i^K, \tag{12}$$

$$W_i^+(t) = N V_i^K. \tag{13}$$

(a) (b)

Fig. 3: The detection results using HOG (a) and YOLOv3 (b). The YOLOv3 can achieve much better result than the HOG from OpenCV library.

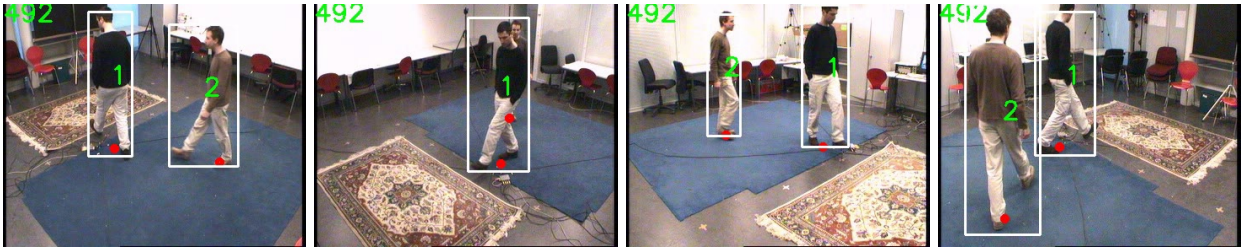
IV. Experiments

To demonstrate the performance of our multi-view multi-target tracking algorithm, we compare our method with other state of the art methods on the public EPFL datasets [8]. The comparison results show that our method can handle occlusion, reidentification and crowd tracking effectively, which can be seen in Fig. 4, Fig. 5 and Fig. 6. The parameters used in the algorithm are set as $\lambda = 0.9, Q = diag\{10, 10, 10, 10, 1, 1, 1, 1\}, R = diag\{0.02, 0.02, 0.02, 0.02\}$ where $diag$ means the diagonal matrix.

A. Compare with MTIC

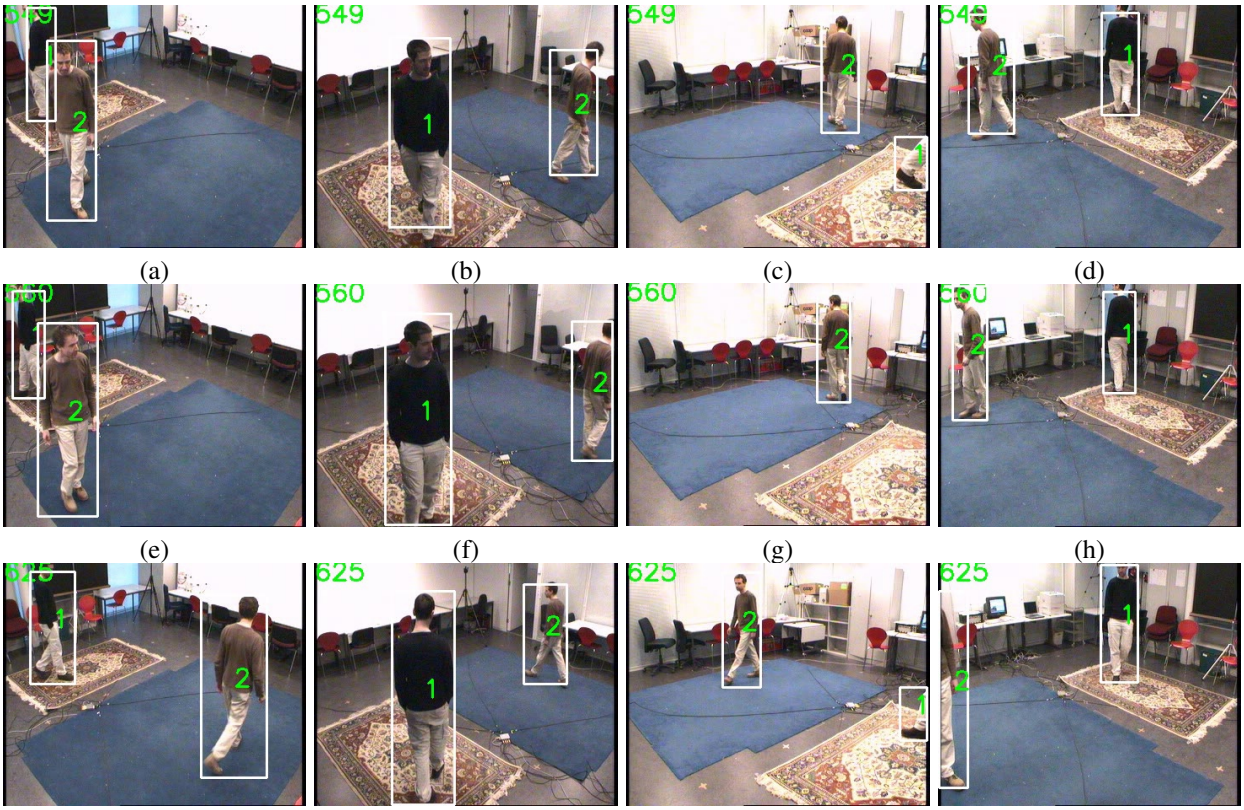
In this section, we compare our method with the MTIC [12], which is a state of the art distributed multi-target tracking method. In contrast to our method, MTIC only uses position information for data association, and has no track process module for ID management. The original MTIC uses histogram of oriented gradient (HOG) to detect humans, which has much worse performance than the YOLOv3 as shown in Fig. 3.

Therefore, we use YOLOv3 for target detection in MTIC to improve its performance. The errors of the estimated trajectories using Laboratory sequences are shown in Fig. 7, where we can see the MTIC only can handle fixed number of



(a) (b) (c) (d)

Fig. 4: Our multi-view multi-target tracking method can handle occlusion problem effectively. The occluded person in (b) can be tracked continuously by fusing multiple view information, where the red dots on the pictures mean the position of the human on the ground plane. The subfigures (a), (b), (c) and (d) are the 492th frames of the video sequences captured from four individual cameras respectively.

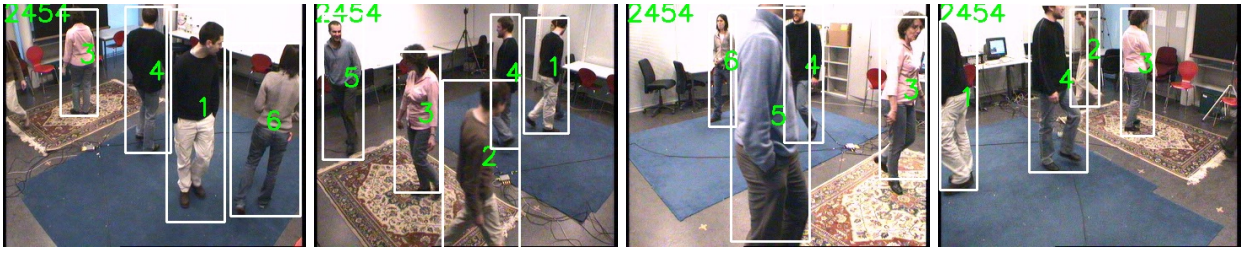


(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Fig. 5: Our multi-view multi-target tracking method can handle reidentification problem effectively. (a), (b), (c) and (d) are the 549th frames, (e), (f), (g) and (h) are the 560th frames, and (i), (j), (k) and (l) are the 625th frames from four cameras. One person is going out of the field of view in (c), totally disappears in (g), and returns back in (k). We can see that the ID of the target is still 1, which proves that our method successfully reidentifies the person.



(a) (b) (c) (d)

Fig. 6: Our multi-view multi-target tracking method can handle crowd tracking effectively.

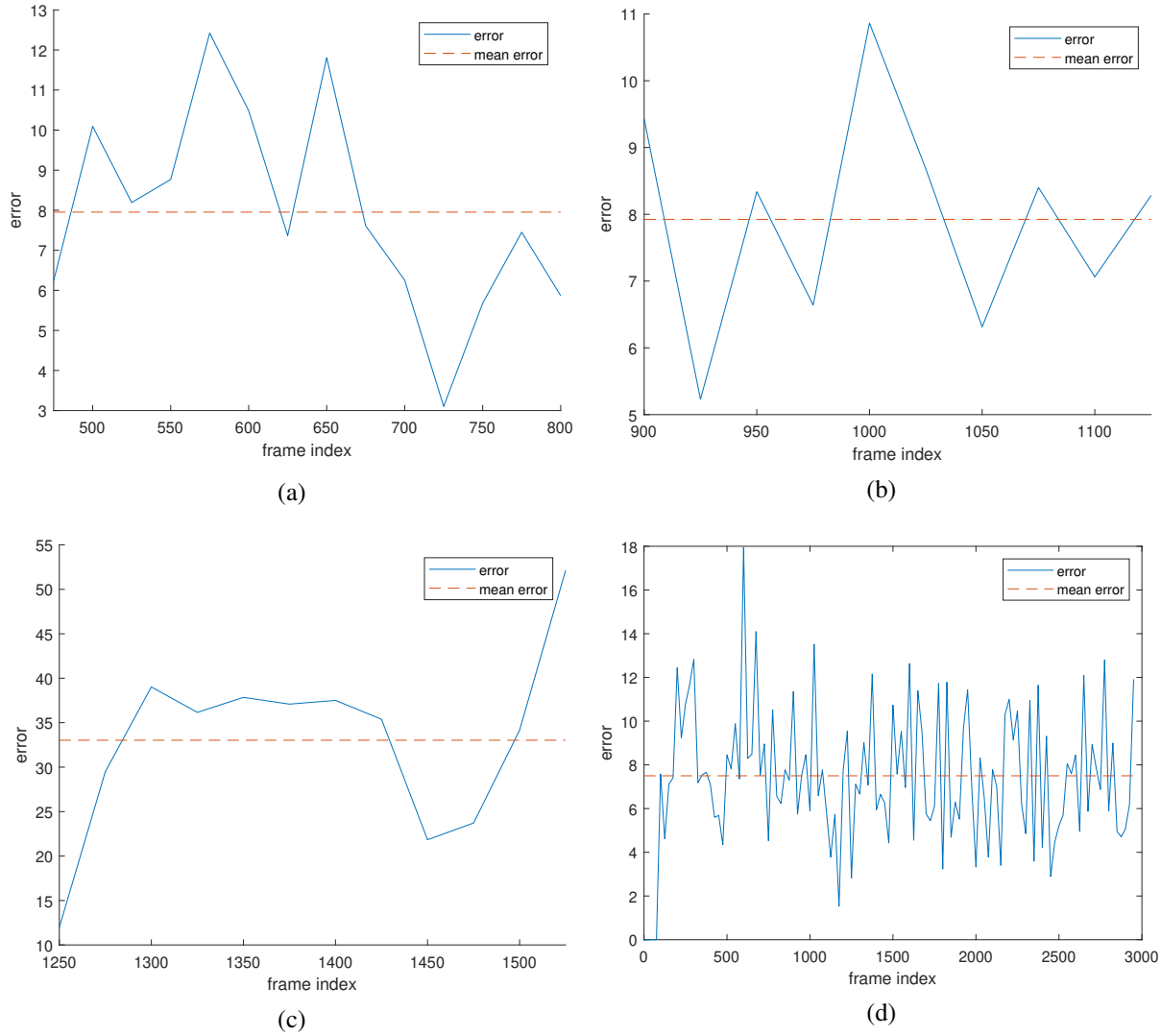


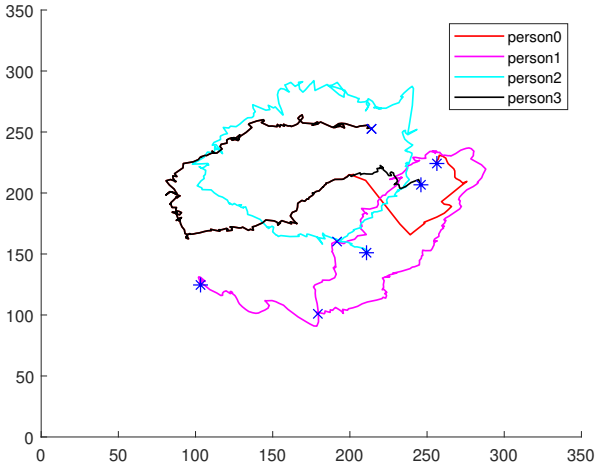
Fig. 7: The MTIC can not handle the situation that the number of targets changes, whereas our ID management method can solve such a problem. Therefore, we manually divide the whole sequence into three sub-sequences for MTIC. (a), (b) and (c) are the trajectory errors of the MTIC algorithm for tracking 2, 3 and 4 persons respectively. (d) is our result for tracking the whole sequence, since the proposed method can create new tracks automatically.

target tracking problem, and get diverged results for tracking four persons. The reason behind this can be seen in Fig. 8, which shows the trajectories of person 0 and person 3 merged together since the MTIC can not distinguish two persons when they are close. In contrast, our method can track the targets successfully since both visual appearance feature and fused spatial location information are used for data association. Further more, our method can handle the problem of varying number of targets using the effective ID management method, which is a necessary function for practical applications.

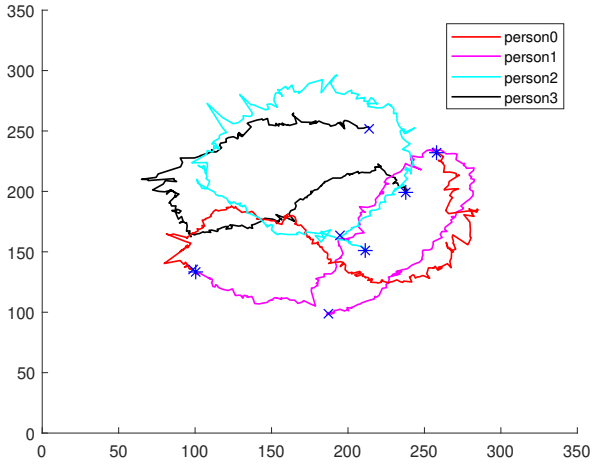
B. Compare with Single View Deep SORT

Deep SORT is a state of the art real time multi-target tracking algorithm using deep features. In contrast to the Deep SORT which uses Faster-RCNN for object detection, we employ YOLOv3 which has comparable performance but with faster processing time [1]. Furthermore, we use a distributed information weighted consensus filter to estimate the position information of targets, which can improve the tracking accuracy for occlusion problems. The quantitative comparison results on the Laboratory sequences can be seen in Table I, which shows our distributed solution has better performance than the original single view Deep SORT. We get the ground truth using an annotation tool provided by [9], which outputs a bounding box and ID for each moving target. The comparison metrics are defined as:

- IDF1: ID F1 score. The ratio of correctly identified targets over the average number of ground-truth and computed targets [16].
- MOTA: multiple object tracking accuracy considers false positives, missed targets and identity switches [17].
- MOTP: multiple object tracking precision considers not only the misalignment between the annotated and the predicted bounding boxes, but also the contribution of



(a) MTIC



(b) Our method

Fig. 8: The performance comparison between our method and the MTIC. Here we track four persons whose trajectories have different colors in the plot. We use '*' and 'x' to depict the start point and end point of the trajectory respectively. We can see the trajectories of person 0 and person 3 are merged together using MTIC in subfigure (a) when they get close. In contrast, our method can track four persons successfully for all frames in subfigure (b).

split and merged tracks [17].

The metric values in Table I are averaged on results from four camera views. We achieve much better performance on the metric IDF1, since IDF1 is mainly affected by the correctness of ID assignment. In contrast to the single view Deep SORT, Our method fuses multi-view information that can better handle occlusion situation and assign right track ID to the target. For efficiency, Deep SORT processes one frame from single view at a cost of about $5ms$, while our method uses $7ms$ to fuse multiple view information using consensus iteration steps.

C. Compare with Other Multi-View Tracking Algorithms

We also compare our method with other state of the art multi-view tracking algorithms on the challenging Terrace and Passageway sequences of EPFL dataset, i.e., POM [7], KSP [8], HTC [9] and PTPE [10]. Both of Terrace and Passageway sequences are recorded in outdoor environment. The challenges of Terrace sequences are the increased number of people, frequent of occlusion and out of field, whereas the Passageway sequences have very poor light condition. For comparison, we use not only the MOTA and MOTP, but also MODA (Multiple Object Detection Accuracy) and MODP (Multiple Object Detection Precision) as metrics, which are defined as:

- MODA: multiple object detection accuracy combines two error sources: false positives, missed targets [17].
- MODP: multiple object detection precision uses spatial overlap information between the ground truth and the system output [17].

The comparison results in Table II show that we achieve the best performance for these challenging scenarios, since out method employs excellent object detector, deep learning features, consensus fusion and efficient ID management.

TABLE I: Compare with single view Deep SORT using the Laboratory sequences

Method	IDF1(%)	MOTA(%)	MOTP(%)
Ours	**89.4**	**80.6**	**84.1**
Deep-SORT	31.0	78.8	83.1

TABLE II: Compare with multi-view algorithms using Terrace and Passageway sequences

Sequence	Method	MODA(%)	MODP(%)	MOTA(%)	MOTP(%)
Terrace	POM	58.5	63.5	57.5	62.6
	KSP	68.3	58.1	65.7	58.3
	HTC	72.6	71.8	72.3	71.6
	Ours	**81.1**	**88.2**	**79.9**	**87.5**
Passageway	POM	32.6	62.5	32.6	60.9
	KSP	40.5	58.9	40.5	57.2
	HTC	43.8	67.1	43.8	67.1
	PTPE	61.0	73.1	60.2	72.2
	Ours	**66.3**	**91.4**	**62.2**	**90.6**

V. Conclusion

In this paper, we propose a new distributed multi-view multi-target tracking framework by combining the advanced deep learning method and distributed information fusion method. The proposed method is simple and direct, and can handle the occlusion and reidentification problems effectively. We demonstrate our idea using the public EPFL datasets, which shows superior results compared to the state of the art methods, such as MTIC, Deep SORT and KSP. We believe that our distributed solution has a great potential for many real world applications due to its scalability and robustness, e.g., large area surveillance. In future, we will extend our work to track 3D models of multiple targets using a 3D camera network.

References

[1] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[2] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, Sep. 2017.

[3] A. T. Kamal, J. A. Farrell, and A. K. Roy-Chowdhury. Information weighted consensus. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 2732–2737, Dec 2012.

[4] S. H. Rezatofighi, A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid. Joint probabilistic data association revisited. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3047–3055, Dec 2015.

[5] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg. Multiple hypothesis tracking revisited. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4696–4704, Dec 2015.

[6] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, Sep. 2016.

[7] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, Feb 2008.

[8] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1806–1819, Sep. 2011.

[9] Y. Xu, X. Liu, Y. Liu, and S. Zhu. Multi-view people tracking via hierarchical trajectory composition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4256–4265, June 2016.

[10] Z. Tang, R. Gu, and J. Hwang. Joint multi-view people tracking and pose estimation for 3d scene reconstruction. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, July 2018.

[11] W. He, W. Tao, K. Sun, L. Xu, Q. Fu, and H. Zhao. Multi-camera object tracking via deep metric learning. In *Sixth International Conference on Optical and Photonic Engineering (icOPEN 2018)*, volume 10827, page 108271I, 2018.

[12] A. T. Kamal, J. H. Bappy, J. A. Farrell, and A. K. Roy-Chowdhury. Distributed multi-target tracking and data association in vision networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1397–1410, July 2016.

[13] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *CoRR*, abs/1905.05055, 2019.

[14] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017.

[15] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics*, 52(1):7C21, 2010.

[16] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set formulti-target, multi-camera tracking. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 17–35, Cham, 2016. Springer International Publishing.

[17] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):319–336, Feb 2009.